GAIA: At the Frontiers of Astrometry
C. Turon, F. Meynadier and F. Arenou (eds)
EAS Publications Series, **45** *(2010) 173-178*
www.eas.org

CLASSIFICATION OF OPTICAL TRANSIENTS: EXPERIENCES FROM PQ AND CRTS SURVEYS

A.A. Mahabal[1], S.G. Djorgovski[1], C. Donalek[1], A.J. Drake[1], M.J. Graham[1], R.D. Williams[1], B. Moghaddam[2] and M. Turmon[2]

Abstract. Synoptic sky surveys are opening up exciting opportunities in time domain astronomy. Gaia will make a great contribution to this field. A crucial factor for good scientific returns is real-time classification of transients, in order to optimize their follow-up. We have been developing infrastructure towards this end starting from the completed Palomar-Quest (PQ) survey, and the ongoing Catalina Real-Time Transient Survey (CRTS). CRTS has been consistently producing transients for almost three years now. We describe here the efforts related to transient classification and event dissemination. Many of the technologies and methodologies we are developing may benefit Gaia.

1 Introduction

A new generation of synoptic sky surveys is opening up time domain as an exciting area of research, touching on essentially every field of astronomy: from the Solar system, through stellar evolution, Galactic structure, extreme relativistic phenomena, and cosmology. These surveys generate massive data streams, starting from a fraction of a TB per night today, and rapidly moving into the Petascale regime, with tens or even hundreds of TB per day (*e.g.* LSST and SKA). Synoptic surveys may be optimized for a particular kind of science, but their data streams can feed multiple scientific investigations.

Detections of transient events are especially interesting. Their ephemeral nature and the need for a rapid follow-up of interesting cases imply a need for a rapid dissemination and classification of events. Today, we are typically dealing with tens of events per night, but in the LSST era the numbers may be $\sim 10^5$/night. Since the follow-up resources are scarce, a rapid identification of the most interesting events is essential.

[1] Caltech, MC 249-17, Pasadena, CA 91125, USA
[2] Jet Propulsion Laboratory, 4800 Oak Grove Dr., Pasadena, CA 91109, USA

DOI: 10.1051/eas/1045030

Article published by EDP Sciences

Many surveys in the near future, both ground-based and from space intend to release transient information in near real time (Gaia, AstroSat, LSST, SKA to name a few). Here we describe the efforts and experiences gained in the course of two ground-based synoptic sky surveys, Palomar-Quest (PQ; `http://palquest.org`; Djorgovski *et al.* 2008), and the Catalina Real-Time Transient Survey, described below. We believe that many of the methodologies developed in the course of this work may be useful for the future surveys, including Gaia.

2 Catalina Real-Time Transient Survey (CRTS)

Catalina Real-Time Transient Survey (CRTS; `http://crts.caltech.edu`) analyses data streams from 3 survey telescopes: The 0.7 m Catalina Sky Survey (CSS) Schmidt Telescope, and the 1.5 m Mt. Lemmon Survey (MLS) Telescope in Arizona, and the 0.5 m Siding Spring Survey (SSS) Schmidt Telescope in Australia. The original goal of these surveys is to look for Earth-crossing asteroids. The cadence is to take four images of the same part of the sky $\sim$ 10 minutes apart, and do so repeatedly as the currently observable sky is covered, adding to several tens to a few hundred revisits over time. We use the object catalogues obtained during the asteroid finding process to look for another kind of transients: those that have brightened up considerably as compared to archival images of the same part of the sky. The surveys are conducted with an unfiltered CCD. Basic calibrations are carried out using field stars and standard astrometry corrections are applied. More details can be found in Drake *et al.* (2009). The points we want to emphasize here are: (1) a large number of epochs are available per pointing leading to some automatic self-follow-up, and (2) even for newly detected transients a series of measurements (several tens) from the past are available, if only as upper limits or with large error-bars. Between the three surveys, nearly three-fourths of the sky is covered, with mainly the Galactic plane excluded.

CRTS is the first fully open survey to publicly release the discovery data on transients as well as the follow-up data that help secure classifications. The transients that are detected are published online in real time using the VOEvent protocol (Williams & Seaman 2007) and are thus freely available for follow-up by anyone listening electronically to the alerts. Over 2000 transients have been detected by CRTS, including more supernovae than any other survey. A large number of ATels, CBETs as well as a few papers have already resulted from this (*e.g.* the brightest SN known so far, see Drake *et al.* 2010).

3 Classification of the transients

Before classifying transients, artifacts masquerading as transients must be removed. Many problems with the data can appear as spurious transients, and in a massive data stream, this is practically inevitable. These can be removed using a supervised clustering method, and a training data set classified by experts "by eye." In PQ we used an ANN-based classifier for this purpose. For each

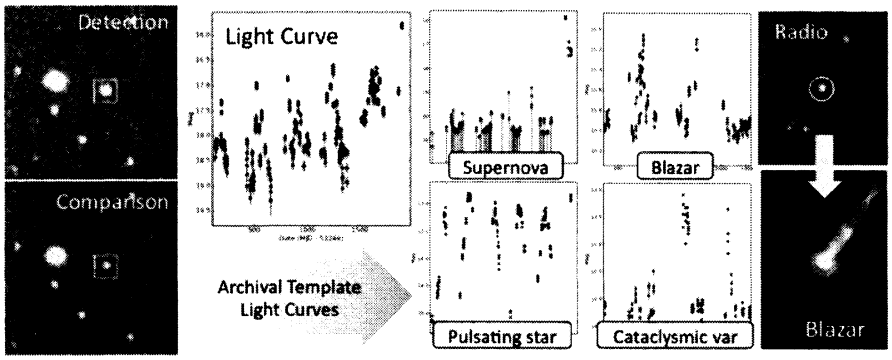

Fig. 1. An example of a human event classification informed by a modest domain expertise and external data: a transient is found with a counterpart visible on the comparison image. Its light curve is extracted from the archival survey images, and compared to the template light curves for typical events; supernova and pulsating star interpretations can be rejected, but there is an ambiguity between the blazar and cataclysmic variable interpretation. Looking at a radio image of the field shows a prominent radio source at the location of the transient; this is fully consistent with the blazar interpretation, and inconsistent with the cataclysmic variable interpretation, thus resolving the issue.

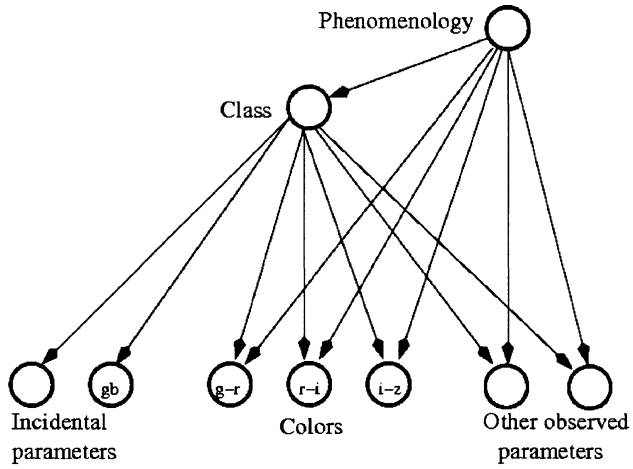

Fig. 2. Architecture of the Naive Bayesian Network currently in use for CRTS transient classification. Each of the three colours is a multinomial node. The output class is also a multinomial node with possible outputs being SN, CV, blazar, AGN, UV-Ceti and all else combined in to a sixth class. Other multinomial nodes that form the input are context parameters like Galactic latitude (gb), proximity to a radio source etc. In the presence of even a single input parameter, output is available. Updating the network is easy when more data become available (both, improved classification and retraining).

object the classifier takes as input a set of measured morphological parameters and returns the probability of it being a real object (Donalek *et al.* 2008).

History and context: For the authenticated transients very little is initially known about it that can help in classifying it. The history is typically made of either faint detections or just upper limits. We routinely cross-match the transient positions to archival catalogues at other wavelengths as well as past surveys like DPOSS, NEAT, SDSS, PQ, etc. These provide a context that can start illuminating the nature of the source better. Presence of a galaxy nearby is a likely indication of the transient being a supernova, whereas a radio source nearby makes it a likely blazar (see Fig. 1).

Bayesian Networks: Such measures, however, need to be quantified. We achieve this using the Bayesian formalism which has the advantage that it can deal with missing data elegantly. Thus, for each parameter, *e.g.* distance to the nearest galaxy, a dataset for past transients is collected. For each previously classified transient type we then have a probability based on the distance. This forms our prior. For every new transient it is then easy to calculate the probability of that object belonging to each of the n classes. With every added observation the priors and the reliability improve. Only near-certain classifications are incorporated into priors. Our current prior size is several hundred and is made of objects classified by experts.

Follow-up Observations like these that can be obtained from archives can be considered to be passive follow-up. We also obtain active follow-up in terms of new images (and when possible spectra) from other telescopes (*e.g.* Palomar 1.5 m, IUCAA Girawali Observatory 2 m telescope etc.) in different filters. These provide colour information as well as additional points for the lightcurve. Priors and probability distributions are formed for the colours as well and used in the Bayesian Network as described above. Figure 2 shows the architecture of the Bayesian network being used to classify objects using the priors as discussed above. The colours used are from follow-up images from the Palomar 1.5 m telescope. Gaia will have the advantage of being able to use the *G*, *BP*, *RP* images to form its own colours to start with.

Owing to the nature of the survey we find large number of SNe and CV, followed by blazars and a few AGN. Our current networks are tuned to classifying objects into these classes as well as smaller ones like UV-Ceti. The rest get collected into a generic "Rest" class. As more observations gather, larger number of subclassifications will be possible. Confusion matrices are generated for known types to study any misclassifications.

Gaussian Process Regression: The cross-matches, proximity parameters, colours etc. form a large but sparsely populated parameter vector for each transient. Follow-up observations may not always be possible, and sometimes the transient location lies outside the footprint of many archives. Under such circumstances a basic characterization of the source may be all that is possible. Gaussian Process Regression (GPR) can be used for some of the classes for possible characterization. These use parameters which are optimized for a given class using known examples and then based on a small number of observations one can state

the likelihood of the transient belonging to each of the classes for which templates have been made. This method is more effective for non-periodic variables (and thus for many of the types we are interested in, *e.g.* SNe, CVs and blazars). As more number of points accumulate, the classification can become more secure. More details can be found in Mahabal *et al.* (2008a, 2008b).

Other techniques: Many more classification methods are being implemented including such diverse methods as Markov Logic Networks (MLN) which allow declarative domain knowledge to be expressed with real-valued weight indicating the strength of statements, Self Organizing Maps (SOM), Support Vector Machines (SVM) and various time-series methods that can deal with large gaps between successive observations. These are being collected in an interoperable way and consistent with Virtual Observatory standards under the DAta Mining and Exploration (DAME) initiative (`http://voneural.na.infn.it`).

Classification Fusion: Different methods use different parts of the parameter vector as input. Sometimes the resulting classifications can seem inconsistent. There has to be a conciliatory way of resolving this. Towards this end we are developing a classifier framework based on a sleeping expert schema (Blum *et al.* 2007) and a fusion module: each specialist makes a prediction only when the instance to be predicted falls within their area of expertise and the results are then combined through the fusion module to have the final probabilistic classification. External and contextual information is used to put experts to sleep or awaken them and to modify online the weight associated with each classifier. The end-result is an improved classification.

4 Event dissemination

As more and more transients compete for scarce follow-up time it becomes even more important to get as much relevant information out there as possible and quickly. Towards this end we make all our transient detection data public so that anyone interested in the transient can quickly get more observations. We also provide early classification information so that groups interested in specific science can choose to follow only those kinds of objects. It is also possible to subselect on parameters like magnitude so that small telescopes do not end up trying to follow impossibly faint targets. All this information is available in multiple formats accessible by clients in various programming languages (as well as in the WorldWide Telescope and Google Sky interfaces) under the SkyAlert (`http://www.skyalert.org`) umbrella (see Williams *et al.* 2009 for details).

For each transient a portfolio is formed where both passive and active follow-up can get added in a transparent manner. It is easy for anyone to add new observations that will help everybody. The classifications too can gradually improve as more data come in. Annotations by experts and citizen scientists are also possible. Semantic harvesting of the comments will be carried out on these. Separate campaigns to harvest Citizen Scientists' abilities to pick patterns (which are not easy to program) are also being planned. The results will be compiled to develop advanced discrimination tools.

5 Relevance to Gaia

In its five year mission Gaia will observe one billion stars an average of 80 times covering most of the sky. Besides a broad band (G), two simultaneous spectra (BP and RP) will be obtained. While the astrometry data will be slow in accumulation, the photometry will allow for quick transient detection and dissemination. The two spectra initially will be just two additional magnitudes. So in a sense the observations are like three real-time colours, something CRTS needs to do separately. Thus the Bayesian Networks equivalent to those being used by CRTS can be used without any change to the architecture. One difference as regards to the history of a transient is that as a spacecraft policy Gaia will transmit data only on detections and for locations specially stipulated for observations. What that means is that for new transients no Gaia-specific history will be available (from that point on it can be added to the list of special locations so that even if the transient fades below certain threshold, forced photometry will continue to be available). For transients with no internal history, data from similar surveys (*e.g.* CRTS) will be very useful.

The current event dissemination architecture can be used without any changes. In the current architecture each new source of transients becomes a stream of events, so that there can be a new Gaia stream. Familiarity of the community and the rich, extensible structure will translate into easy adoption and productive science.

We are grateful to the staff of Palomar and Keck Observatories for their help, and to our collaborators in the PQ and CRTS teams. This work was supported in part by the NSF grants AST-0407448, AST-0909182, NSF-0915473, NASA grant 08-AISR08-0085, by Microsoft Research, and by the Ajax Foundation.

References

Blum, A., & Mansour, Y., 2007, J. Mach. Learn. Res., 8, 1307

Djorgovski, S.G., *et al.*, 2008, Astron. Nach., 329, 263

Donalek, C., *et al.*, 2008, in "Proc. Intl. Conf. on Classification and Discovery in Large Astronomical Surveys", AIPC, 1082, 252

Drake, A.J., *et al.*, 2009, ApJ, 696, 870

Drake, A.J., *et al.*, 2010, ApJ, 718, L127

Mahabal, A.A., *et al.*, 2008, "Proc. Intl. Conf. on Classification and Discovery in Large Astronomical Surveys", AIPC, 1082, 287

Mahabal, A.A., *et al.*, 2008, Astron. Nach., 329, 288

Williams, R.D., *et al.*, 2009, in "Proc. ADASS XVIII", ed. D.A. Bohlender, D. Durand, & Patrick Dowler, ASPC, 411, 115

Williams, R.D., & Seaman, R.L., 2007, in "NVO Tools and Techniques", ed. M.J. Graham, M.J. Fitzpatrick, & T.A. McGlynn, ASPC, 383, 425